A Multi Cue Discriminative Approach to Semantic Place Classification

Marco Fornoni, Jesus Martinez-Gomez, and Barbara Caputo**

Idiap Research Institute
Centre Du Parc, Rue Marconi 19
P.O. Box 592, CH-1920 Martigny, Switzerland
{mfornoni, jmartinez, bcaputo} @idiap.ch

Abstract. This paper describes the participation of Idiap-MULTI to the Robot Vision Task at imageCLEF 2010. Our approach was based on a discriminative classification algorithm using multiple cues. Specifically, we used an SVM and combined up to four different histogram-based features with the kernel averaging method. We considered as output of the classifier, for each frame, the label and its associated margin, which we took as a measure of the confidence of the decision. If the margin value is below a threshold, determined via cross-validation during training, the classifier abstains from assigning a label to the incoming frame. This method was submitted to the obligatory task, obtaining a maximum score of up to 662, which ranked second in the overall competition. We then extended this algorithm for the optional task, where it is possible to exploit the temporal continuity of the sequence. We implemented a door detector so to infer when the robot has entered a new room. Then, we designed a stability estimation algorithm for determining the label of the room where the robot has entered, and we used this knowledge as a prior for the upcoming frames. Our approach obtained a score of up to 2052 in the obligatory task, ranking first.

1 Introduction

This paper describes the algorithms used by the Idiap-MULTI team at the third edition of the Robot Vision task, held under the umbrella of the ImageCLEF 2010 evaluation challenge. The focus of the Robot Vision task has been, since its first edition in 2009, semantic place localization for mobile robots, using visual information. This year, the task posed two distinctive research questions to participants: (1) can we design visual recognition algorithms able to recognize room categories, and (2) can we equip robots with methods for detecting unknown rooms?

The Idiap-MULTI team took a multi cue discriminative approach for addressing both issues. The core of our methods, both in the obligatory and optional tasks, is an SVm classifier, trained on a large number of visual features, combined

** This work was supported by the SNSF project MULTI (M. F.) and by the Spanish "Junta de Comunidades de Castilla-La Mancha" (PCI08-0048-8577 and PBI-0210-7127 Projects, J. M.-G.).

together via a flat average of kernels [Gheler09]. This outputs, for each frame of the testing sequence, a classification label and a measure of the confidence in the decision. These two informations are then used to evaluate if the perceived room is one of those already seen, or if it is unknown to the system. Figure 1 gives an overall overview of the training and classification steps.

In the rest of the paper we provide a detailed description of each step outlined in the diagrams: section 2 gives an overview of the oversampling strategy, devised to increase robustness. Section 3 describes the features used, and section 4 the cue integration approach. Section 5 and 6 describes into details the algorithms used for the obligatory and optional tasks. We report the experimental results in section 7. The paper concludes with an overall discussion.

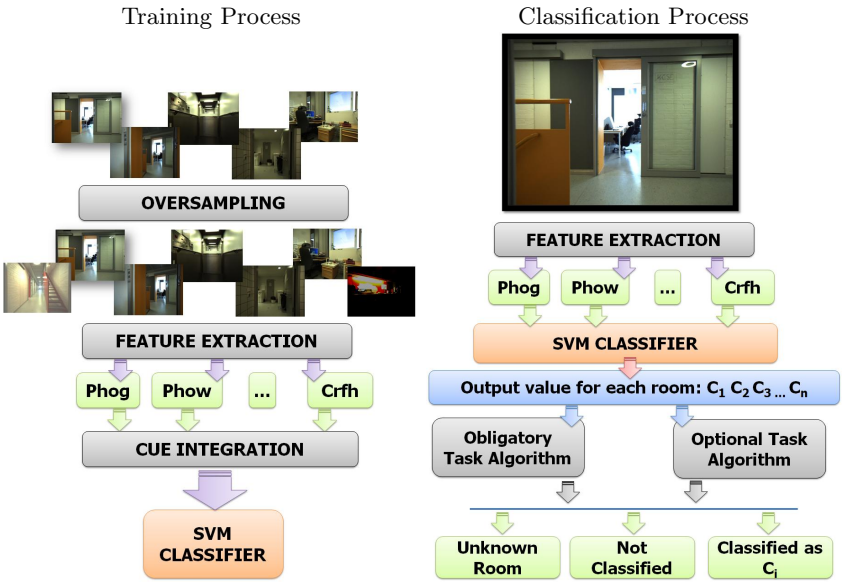

Fig. 1. Overall overview of the training and the classification process

2 The Oversampling Strategy

The capability to recognize room categories implies robustness to slight variations in the rooms' appearance. To achieve this, we propose, as a pre-processing step, to increase the number of training frames by applying simulated illumination changes to the original training frames. We generate new frames using the original training frames as templates. We apply colour modifications to that templates, trying to emulate the effect of extreme low/high lighting environments (figure 2). We increased the original training set adding an additional sequence (the size was 30% of the original training sequence) with images generated increasing or decreasing the luminance component for all the pixels. Even though in principle categorical variations are different from lighting variations, preliminary experiments indicate that this pre-processing step is beneficial.

Fig. 2. Example of the oversampling process, where two new training frames are generated using a original training image

3 Feature Extraction

As features, we chose a variety of global descriptors representing different features of the images. We opted for histogram-based global features, mostly in the spatial-pyramid scheme introduced in [**?**]. This representation scheme was chosen because it combines the structural and statistical approaches: it takes into account the spatial distribution of features over an image, while the local distribution is in turn estimated by mean of histograms; moreover it has proven to be more versatile and to achieve higher accuracies in our experiments.

The descriptors we have opted to extract belong to five different families: Pyramid Histogram of Orientated Gradients (PHOG) [**?**], Sift-based Pyramid Histogram Of visual Words (PHOW) [**?**], Pyramid histogram of Local Binary Patterns (PLBP) [**?**], Self-Similarity-based PHOW (SS-PHOW) [**?**], and Compose Receptive Field Histogram (CRFH) [**?**]. Among all these descriptors, CRFH is the only one which is not computed pyramidly. For the remaining families we have extracted an image descriptor for every value of $L = \{0, 1, 2, 3\}$, so that the total number of descriptors extracted per image is equal to 25 (4+4 PHOG, 4+4 PHOW, 4 PLBP, 4 SS-PHOW, 1 CRFH). The exact settings for each descriptor are summarized in Table 1.

DESCRIPTOR	SETTINGS	L
PHOG_{180}	range= [0, 180] and $K = 20$	$\{0, 1, 2, 3\}$
PHOG_{360}	range= [0, 360] and $K = 40$	$\{0, 1, 2, 3\}$
$\text{PHOW}_{\text{gray}}$	$M = 10$, $V = 300$ and $r = \{4, 8, 12, 16\}$	$\{0, 1, 2, 3\}$
$\text{PHOW}_{\text{color}}$	$M = 10$, $V = 300$ and $r = \{4, 8, 12, 16\}$	$\{0, 1, 2, 3\}$
$\text{PLBP}_{8,1}^{riu2}$	$P = 8$, $R = 1$, RotationInvariantUniform2 version	$\{0, 1, 2, 3\}$
SS-PHOW	$M = 5$, $V = 300$, $S = 5$,$R = 40$, $nRad = 4$ and $nTheta = 20$	$\{0, 1, 2, 3\}$
CRFH	Gaussian-Derivatives=$\{L_x, L_y\}$, $K = 14$ and $s = \{1, 2, 4, 8\}$	

Table 1. Settings of the image descriptors

4 Cue Integration

Categorization is a difficult task and this is particularly true for indoor visual place categorization. Indoor environments are indeed characterized by an high variability in the visual appearance within each category, mainly due to clutter, occlusion, partial visibility of the rooms and local illumination changes. To further complicate matters, a robot is supposed to interact responsively with its environment and therefore a strong requirement is efficiency. We decided to combine these two issues by using a very efficient cue integration scheme, namely kernel averaging [**?**,**?**]. Our approach consists of two steps:

1. pre-select the visual cues which are found to maximize the performances, when integrated together.
2. compute the average-kernel over the preselected features as an effective and efficient cue-integration method

In order to select the best visual cues to be combined together we have performed a pre-selection of the corresponding pre-computed kernel matrices, by using a simple depth-first search on the tree of all possible combinations of features. Since the number of all possible combinations of n features is $2^n - 1$, we have adopted the following efficiency measures to make the computation feasible:

- estimate the accuracy of a given combination of kernels using only a subsample of the training and validation set (10 and 30 percent)
- prune down the tree of all possible combinations, by imposing a condition on the improvement which has to be satisfied in order to explore a branch: if the improvement achieved by averaging the kernel k_2 with the kernel k_1 is less then or equal to a threshold ($ratio * accuracy\_of\_k_1$), the branch is not further explored. However if averaging k_3 with k_1 does satisfy the condition, the branch k_1-k_3-k_2 is explored. Finally if the branch k_1-k_2 has already been explored, obviously the branch k_1-k_3-k_2 is not explored again.

An example exploration is shown in figure 3.

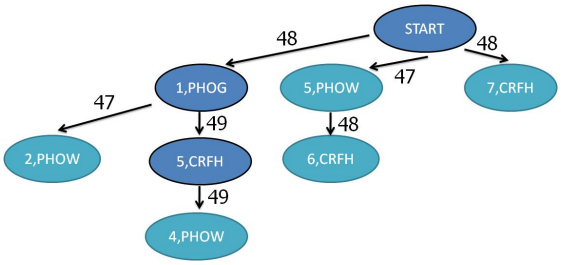

Fig. 3. Search example. Each node represents the average kernel computed with all the features of the path starting from the root node. The number beside the feature name in each node represents the order in which the tree is visited, while the arc's weight represents the accuracy obtained by averaging the successor kernel, to the ancestor one. If this accuracy is not greater then the accuracy on the ancestor arc, the combination is discarded. The best combination retrieved in this example is PHOG+CRFH

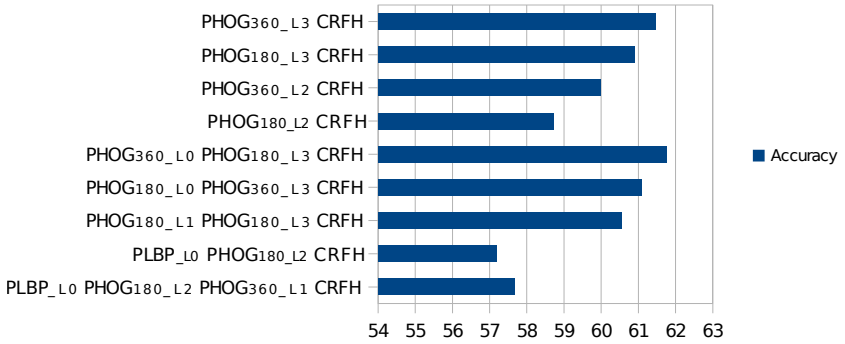

Fig. 4. Performance of best combinations returned by the algorithm (ordered with respect to the number of kernels used), as measured by the accuracy on validation set

Best performing combinations selected are shown in figure 4, where we have sorted them with respect to the number of image descriptors used and we have taken into account only the best combinations with a maximum of four cues. For our final runs we used the following combinations of visual cues:

- PHOG_{360_L3}, CRHF
- PHOG_{180_L3}, CRFH
- PHOG_{360_L0}, PHOG_{180_L3}, CRFH
- PHOG_{180_L0}, PHOG_{360_L3}, CRFH
- PHOG_{180_L1}, PHOG_{180_L3}, CRFH
- PLBP_{L0}, PHOG_{180_L2}, PHOG_{180_L1}, CRFH

which correspond to the two best combinations with 2 cues, the three best combinations with 3 cues and the best combination with 4 cues.

5 Obligatory Task: The Algorithm

For the obligatory task, each test frame has to be classified without taking into account the continuity of the test sequence. Each test frame will be classified just using the SVM after the feature extraction step with the cue integration. Our first approach was just label each test frame with the room (class) that obtained the highest output value, but wrong classifications obtain negative values for the task score (as will be observed in section 7).

Our algorithm post-process the output obtained by the SVM to avoid classifying a test frame if it is not very confidence with the correct class. We normalize the output obtained with the SVM classifier for the test sequence, obtaining (for each test frame) 8 numeric values between -1.0 and $+1.0$ corresponding to each one of the training rooms. A test frame f_i will be labelled with class C_j only when the normalized output value for that class $O_{i,j}$ is above a threshold value and $O_{i,j}$ clearly overcomes all the others output values.

All thresholds were obtained with the preliminary experiments using the validation sequence provided by the task organizers. For these preliminary experiments, we observed that for a big percentage of the validation sequence, just

a class obtained a positive output value. Moreover, large and small office presented as most problematic rooms and Printer Area, Recycle Area and Toiled obtained best accuracy.

For the parameter tuning, we used a classical Hill Climbing algorithm for all thresholds (we have 8 thresholds for each feature combination). A threshold value of 0.0 means that none of the test frames will be classified using that class and 1.0 will be used if we highly trust the classification algorithm for a selected class.

For the Hill Climbing algorithm, we tested positive and negative variations for the threshold values. These variations will be performed if the score obtained for the obtained run with the selected threshold (and using the validation sequence as test sequence) does not decrease. This greedy method has high risks of failing into local optima, and so we perform three executions using 0.25, 0.5 and 0.75 as initial values, selecting as final threshold value that achieving the highest score.

6 Optional Task: The Algorithm

For the optional task, we are allowed to exploit the temporal continuity in the sequence. We therefore implemented a door detector for estimating when the robot moves into a new room. This information, coupled with a stability estimation algorithm, can be useful for classifying a sequence of consecutive test frames. We estimate the stability of the classification process using the last n frames and their associated labels, obtained with the classification algorithm used for the obligatory task. A room is selected as the most probable label for the incoming data if at least the last n frames were classified with that label. This method is used for labeling frames for which the classification algorithm has a low level of confidence and therefore abstains to take a decision. The process is initiated every time the door detector signals that the robot has entered a new room.

Door detection algorithm

We developed a basic door detection algorithm for indoor office environments as those used for the Robot Vision task. When the robot moves from a room to a new one, acquired images show two vertical rectangles with the same colour. The width of both rectangles increases when the robot gets closer to the door. The image processing algorithm consists of a Canny filter [**?**] to extract all the edges of the images. After this step, we use the Hough transform [**?**] for lines detection and we discard all the non vertical lines. Finally, we measure the average colour value between each two vertical lines, removing non homogeneous colour distributions (blobs). All these process can be observed in Fig. 5, where we detect three colour homogeneous blobs (two of them can be used to detect the door crossing)

Once we have extracted all the key blobs from a frame, we have to study the time correspondence for these blobs between this frame and the last frames. If two blobs with the same average colour are increasing for new frames we are reaching a door and both blobs are marked as candidates. Preliminary candidates will be selected as definitive ones if one of the two blobs starts decreasing after reaching the largest size at the left (right) of the image. Figure 6 shows four consecutive training frames, where white rectangles represent blobs, preliminary

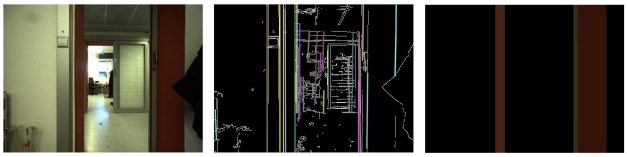

Fig. 5. Complete process to extract blobs. Left: original image. Centre: Vertical lines detection. Right: Homogeneous colour distributions between two vertical lines

candidates are labelled with a P and definitive candidates with a P. Green rectangles for the bottom images represent the time correspondence for each blob in the last frames.

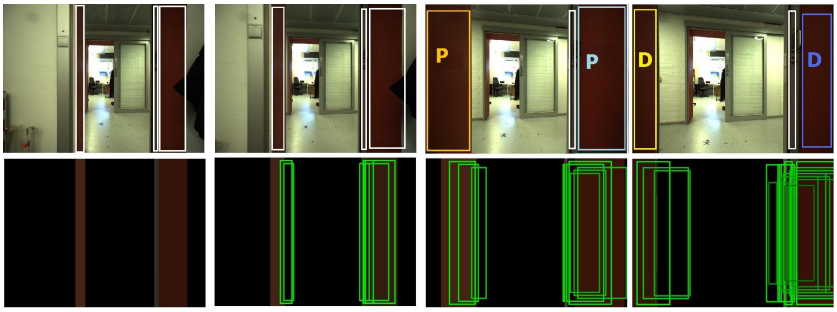

Fig. 6. Door detection for four consecutive frames. Top images are the original frames using P for preliminary candidates and D for definitive ones. Bottom images show the blobs extracted and time correspondence

Additional Processing

For the additional processing we use the output obtained with the algorithm for the obligatory task. For each test frame, that algorithm obtains a class value or leaves it without classifying. Our processing tries to detect two situations: stable process and unstable process. All the internal threshold values were obtained from preliminary experimented developed with the validation sequence as testing set:

- *Stable estimation* The process is stable if, after crossing the last door, most of the frames were preliminary labelled with the same class C_i. In such situation we will use C_i to label test frames not labelled by the classification algorithm. The process will be considered stable when at least the last 18 frames have been classified with the same label.
- *Unstable estimation* Instability will appear when preliminary class values for the last frames is not the same or they were not labelled. If this situation continues for a large number of frames, the process will not be able to achieve a stable situation. We assume that the process is unstable when the number

of frames since we achieved a stable situation is greater than 50. Facing an unstable situation, the additional processing will label with the special label "Unknown" all the frames not labelled previously. This label is used for classifying new rooms not imaged in the training/validation sequences.

7 Experiments

Our algorithms were evaluated following the procedure proposed by the organizers of the RobotVision@ImageCLEF 2010 competition. A training set containing 2741 frames had to be classified using a room label, marked as unknown or not classified. Performance was evaluated according to a pre-defined score function.

7.1 Obligatory Task: Results

We submitted a total of twelve runs to the obligatory task. These runs were divided into two sets, one with parameters determined via cross validation, one without, as described in section 5. Each of the two sets comprises six experiments using the same exact feature combinations.

Rank	Feature Combination	Score	Cross-Validation
2	$PHOG_{180\_L0}$ $PHOG_{360\_L3}$ CRFH	662	✓
3	$PHOG_{360\_L0}$ $PHOG_{180\_L3}$ CRFH	657	
4	$PHOG_{180\_L1}$ $PHOG_{180\_L3}$ CRFH	645	✓
5	$PHOG_{180\_L0}$ $PHOG_{360\_L3}$ CRFH	644	
9	$PHOG_{360\_L3}$ CRFH	637	✓
10	$PHOG_{360\_L0}$ $PHOG_{180\_L3}$ CRFH	636	✓
11	$PHOG_{180\_L1}$ $PHOG_{180\_L3}$ CRFH	629	
12	$PLBP_{\_L0}$ $PHOG_{180\_L2}$ $PHOG_{360\_L1}$ CRFH	628	✓
13	$PHOG_{360\_L3}$ CRFH	620	
14	$PLBP_{\_L0}$ $PHOG_{180\_L2}$ $PHOG_{360\_L1}$ CRFH	612	
15	$PHOG_{180\_L3}$ CRFH	605	
17	$PHOG_{180\_L3}$ CRFH	596	✓

Fig. 7. Ranking of our submitted runs, combination features employed, score and checkmark if the result has been obtained using parameters estimated via cross-validation

Our best score ranked second in the competition, with a difference with respect to the winner of only -2.22%. It was obtained using the cue combination: $PHOG_{180\_L0}$ $PHOG_{360\_L3}$ CRFH, with γ and C estimated via cross-validation. Also our second best score (ranked third) was obtained with a combination of the type: $PHOG_{\_L0}$ $PHOG_{\_L3}$ CRFH, but the quantization of the orientation space was in this case swapped: 360 for the $PHOG_{\_L0}$ and 180 for the $PHOG_{\_L3}$.

As shown in figure 8, most of the experiments where the cross-validation step was performed obtained a higher performance.This improvement is confirmed also by computing the average score in the two sets.

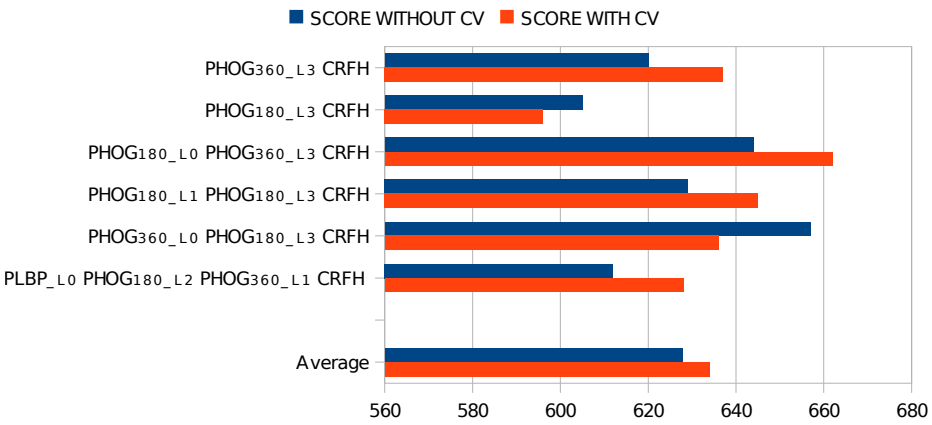

Fig. 8. Score obtained by the our submitted runs. The same experiments have been performed using the parameters obtained by cross-validation and parameters fixed

This can be explained using the parameters values obtained: the γ parameters estimated via cross-validation ($0 < \gamma < 4$) were in general much lower than the corresponding values obtained as the average pairwise χ^2 distance between histograms ($1.5 < \gamma < 16$). The SVM C parameter obtained by the cross-validation ($5 < C < 35$) was also not far from the default value of 10, which is an unusually low value for a classification task (common values are often between 100 and 1000). The low value of the C parameter enforced a stronger regularization of the solution, thus improving the generalization capability of the classifier.

It is important to say, however, that our second best performance on this task was obtained without executing any cross-validation step and nonetheless turned out to outperform the corresponding cross-validated one. Also in this case one explanation for the failure of the cross-validation could be found by looking again at the γ values obtained: one of the three kernels averaged (PHOG_{360_L0}) obtained a very low value (0.000031). With this setting the kernel matrix is almost 1 for most of the couples and when computing the average kernel it only plays the role of a (smoothing) constant. The cross-validation algorithm in this case got stuck in a local minima in which the information added by the PHOG_{360_L0} kernel to the combination was very limited and the final performance was not improved.

7.2 Optional Task: Results

For the optional task we submitted twelve runs, using the same combination of features and cross-validations that for the obligatory task. Fig. 9 shows all the results, where it can be observed that all our runs achieved first positions for this task.

Rank	Feature Combination	Score	Cross-Validation
1	$\text{PLBP}_{\_L0}$ PHOG_{180_L2} PHOG_{360_L1} CRFH	2052	✓
2	PHOG_{180_L3} CRFH	1770	✓
3	PHOG_{180_L0} PHOG_{360_L3} CRFH	1361	✓
4	PHOG_{180_L1} PHOG_{180_L3} CRFH	1284	✓
5	PHOG_{360_L0} PHOG_{180_L3} CRFH	1262	✓
6	PHOG_{180_L1} PHOG_{180_L3} CRFH	1090	
7	PHOG_{180_L0} PHOG_{360_L3} CRFH	1028	
8	PHOG_{360_L0} PHOG_{180_L3} CRFH	1019	
9	PHOG_{360_L3} CRFH	963	✓
10	PHOG_{360_L3} CRFH	916	
11	$\text{PLBP}_{\_L0}$ PHOG_{180_L2} PHOG_{360_L1} CRFH	886	
12	PHOG_{180_L3} CRFH	682	

Fig. 9. Ranking of our submitted runs, combination features employed, score and checkmark if the result has been obtained using parameters estimated via cross-validation

Only other two groups (CAOR and DYNILSIS) submitted runs for this task, and their best scores were consistently smaller than all our runs (62.0 and −67.0 respectively). Therefore, our group was the winner of the optional task. If we compare figures 9 and 7, our algorithm for the optional task allows us to increase the final score for all the feature combinations. This increase proves the goodness of our proposal for exploiting the continuity of the test sequence.

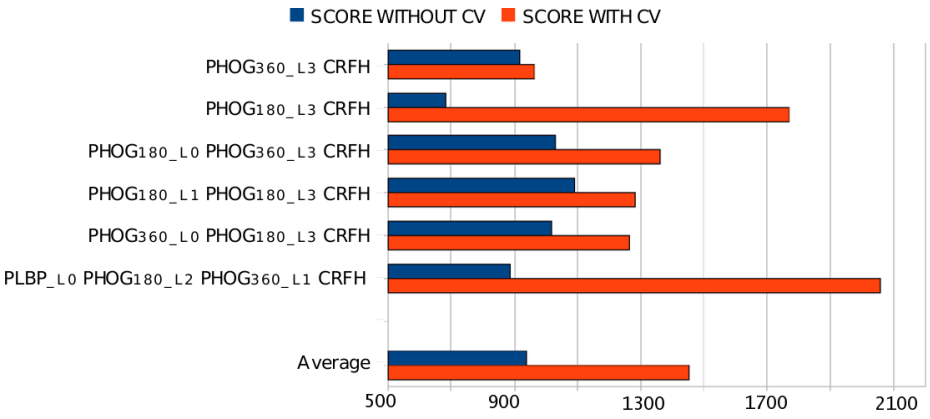

Fig. 10. Score obtained by the our submitted runs. The same experiments have been performed using the parameters obtained by cross-validation and parameters fixed

Fig. 10 shows a complete comparison for the score obtained for each feature combination, with and without using a cross-validation step. It is worth to note

that the final score was always noticeably improved by using the cross-validation step.

8 Conclusions

This paper describes the participation of Idiap-MULTI to the Robot Vision task at ImageCLEF 2010. We participated to both the obligatory and optional tracks with algorithms based on an SVM cue integration approach. Our best runs in the two tracks ranked respectively second (obligatory track) and first (optional track), showing the effectiveness of our approach.